Adaptive Fine-tuning for Transferring a U-net Hydrography Extraction Model using K-means

Lawrence V. Stanislawski, Ethan Shavers, Neal Pastick, Phil Thiem, Shaowen Wang, Nattapon Jaroenchai, Zhe Jiang, Barry Kronenfeld, Barbara P. Buttenfield, Adam Camerer

ABSTRACT: The United States Geological Survey (USGS) coordinates the collection of hydrographic features derived from remotely sensed interferometric synthetic aperture radar (IfSAR) elevation and intensity data in Alaska. Hydrographic features are cartographic representations of surface water features such as stream, rivers, lakes, ponds, canals, etc. Collection and validation procedures involve complex automated and manual techniques that furnish snapshots of hydrographic vector data that exist during the IfSAR surveys. The dynamic nature of fluvial conditions warrants monitoring and updating hydrographic data, but extraction procedures for updates can be cost prohibitive. This paper overviews progress on automated workflows to extract hydrography from IfSAR data using deep learning methods trained and tested with USGS collected hydrography data. This research tests transfer learning methods on a well-performing U-net model trained on a 4600 square kilometer (sq km) base model area in northcentral Alaska. The base model is transferred and fine-tuned to regions in the target domain covering roughly 127,000 sq km. The target domain is subdivided into areas with similar hydrogeomorphic conditions using principal components and k-means clustering, and the base model is adaptively fine-tuned to each hydrogeomorphic class by selecting training watersheds from each cluster within the target domain. Results are compared with transfer learning that is fine-tuned with a random sample of watersheds in the target domain.

KEYWORDS: hydrography, machine learning, transfer learning, adaptive fine-tuning, k-means clustering, Alaska, remote sensing

Introduction

Water on the Earth's surface is a critical resource. Changes to the distribution of surface water can have considerable social, economic, and environmental impacts. The USGS National Hydrography Dataset (NHD) (https://www.usgs.gov/national-hydrography/national-hydrography-dataset) is a vector database of all surface water features in the United States used for land resource management and hydrologic investigations. It includes cartographic representations of features such as streams, rivers, lakes, ponds, and canals compiled at 1:24,000 or larger scale that are displayed on multiple scales of USGS topographic maps. The NHD is being converted to a more efficient hydrologic format in the 3D Hydrography Program (3DHP, https://www.usgs.gov/3DHP) and with updated and more detailed feature representations that are derived from high-resolution elevation data, which includes 1-meter (m) resolution elevation models derived from lidar and, in Alaska, 5-m resolution elevation data derived from IfSAR (https://www.usgs.gov/3d-elevation-program). Collecting, compiling, and validating high-resolution 3DHP data is a complex and costly process that integrates automated and manual techniques to provide single snapshots of vector hydrographic data associated with the date of the remotely sensed information from which it is derived. The dynamic nature of fluvial processes, including seasonal and annual permafrost changes, warrants regular monitoring and updates to hydrographic data, but current collection techniques can be cost prohibitive.

This research demonstrates automated workflows to extract hydrography from IfSAR data in Alaska using deep learning methods trained and tested with the best available USGS hydrography data. Stanislawski et al. (2021) demonstrated using the deep learning U-net neural network model for extracting both linear flow network features (streams, rivers, etc.) and polygonal features (wider streams, lakes, ponds, etc.). Subsequently, they tested methods to transfer the trained U-net model to other areas. They found that predictions refined to nearby base-model domains were more accurate on average than were model predictions refined to domains further away and less similar to the base model domain (Stanislawski et al., 2023). This paper further tests methods to transfer the base U-net model to a larger target domain and classify the domain using k-means clustering. The goal is to determine if the base model can be further refined to the target domain using knowledge about the distribution of hydrogeomorphic conditions and whether such methods are well-suited for hydro-feature collection and update.

Methods

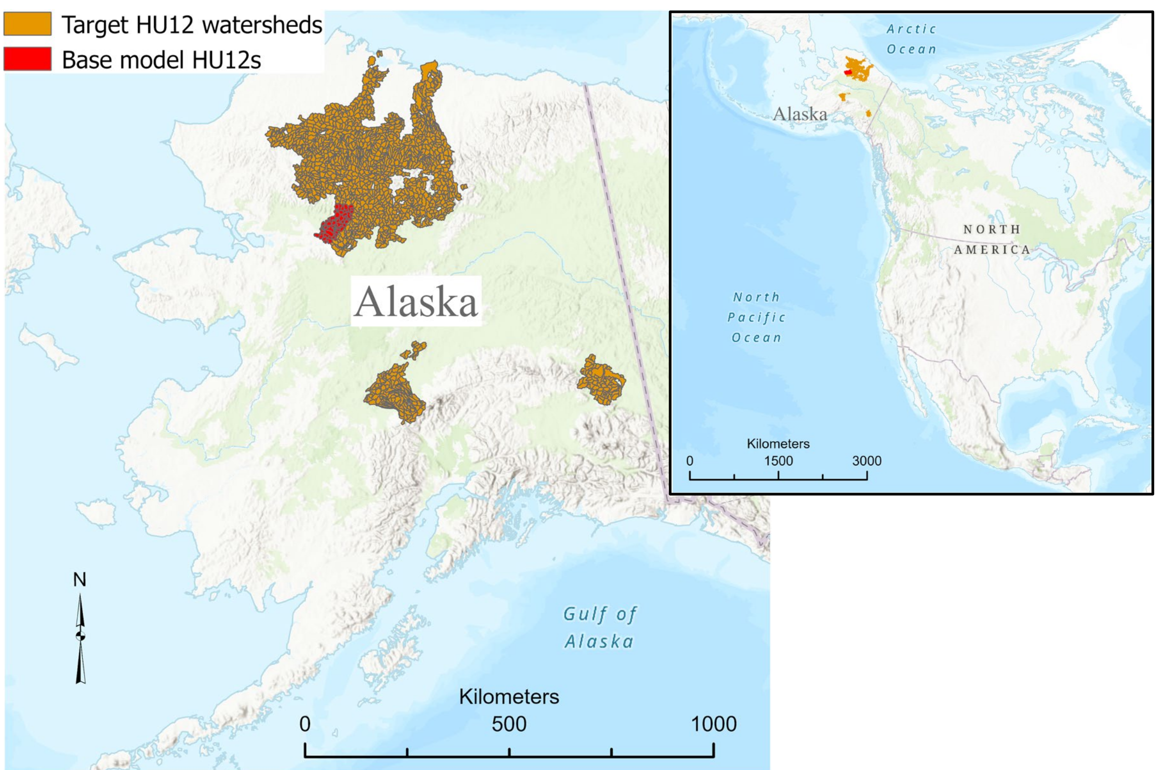

Figure 1: Distribution of 50 12-digit hydrologic unit (HU12) watersheds in the base model area and 1,170 HU12 watersheds in the target domain in Alaska.

The base model area is diverse and covers about 4,600 sq km in northcentral Alaska, including a part of the Kobuk River basin (Figure 1). It is comprised of 50 12-digit hydrologic unit (HU12) watersheds from the USGS Watersheds Boundary Dataset (https://www.usgs.gov/national-hydrography/watershed-boundary-dataset) ranging in size from 30 to 240 sq km. The base U-net model is trained with 12 of the 50 watersheds in the base model area using 3,600 sample windows of 128x128 pixels distributed in each training watershed (i.e., 43,200 total sample windows). U-net is a fully convolutional neural network having a contracting path and an expanding path that combine local detail information from the contracting path with more general data in the expanding path to assemble detailed pixel-wise predictions (Ronneberger et al., 2015). Reference data used for training, validation, and testing consists of 5-m resolution raster versions of the vector hydrography data from the NHD that was derived from 2012-2013 IfSAR data using semi-automated methods (Stanislawski et al., 2021). The raster reference data was filtered to remove pixels likely generated in error, as determined from comparisons with channel depth estimates derived from a 2-D shallow water drainage model (Mitasova et al., 2004). The ten raster input layers used in the U-net model are derived for each HU12 watershed from the IfSAR data and co-registered to 5-m resolution. These raster layers include a digital surface model, two shallow-water drainage models that accentuate variations in surface roughness and diffusion, positive openness, negative openness, orthorectified radar return intensity, sky view factor, curvature, a topographic position index using an 11x11 kernel, and a topographic wetness index (Stanislawski et al., 2021). The base model provides F1 scores, averaging about 80 percent for the 38 test watersheds.

In transfer learning, a model trained on a source domain is refined to a target domain with a relatively small amount of additional training data derived from the target domain. The refined model is applied to the target domain. Pre-training a base model with a portion of the source domain that is similar to the target domain can enhance transfer learning (Cui et al., 2018). However, developing a hydro-feature extraction model that performs well within its domain requires many trials and accurate training data. Therefore, re-using the weights from a good base model in other domains is beneficial wherever possible, which can substantially reduce processing time and training data requirements compared to retraining a model from scratch. In this work, the target study area is subdivided into areas with similar hydrogeomorphic conditions using principal components and k-means clustering. The same base model is adaptively fine-tuned to each hydrogeomorphic class by selecting training watersheds from each cluster in the target domain. Results are compared with transfer learning that is fine-tuned with a random sample of watersheds in the target domain.

The target domain consists of 1,170 HU12 watersheds, mainly to the east and north of the base model area and covers nearly 127,000 sq km (Figure 1). To subdivide the study domain into k-means clusters, the mean, median, and standard deviation are computed for the ten 5-m resolution input raster layers for each HU12 watershed in the base model and target areas (i.e.,1220 HU12 watersheds). Summary statistics for near surface (within 1 m) permafrost probability estimates determined from Landsat 7 Thematic Mapper (2008-2011) and other data at 30-m resolution (Pastick et al., 2015) are included after upsampling to an IfSAR-matching 5-m resolution. Summary statistics for 2019 forest canopy height estimates are also determined for each watershed from the Global Ecosystem Dynamics Investigation lidar instrument and Landsat data at 30-m resolution (Potapov et al., 2021), with upsampling to 5-m resolution. Data noise is reduced in the summary statistics by selecting a set of principal components computed from the 12 layers of summary statistics for all 1,220 HU12 watersheds in the study area. The selected principal components explain more than 90 percent of the variation in the data.

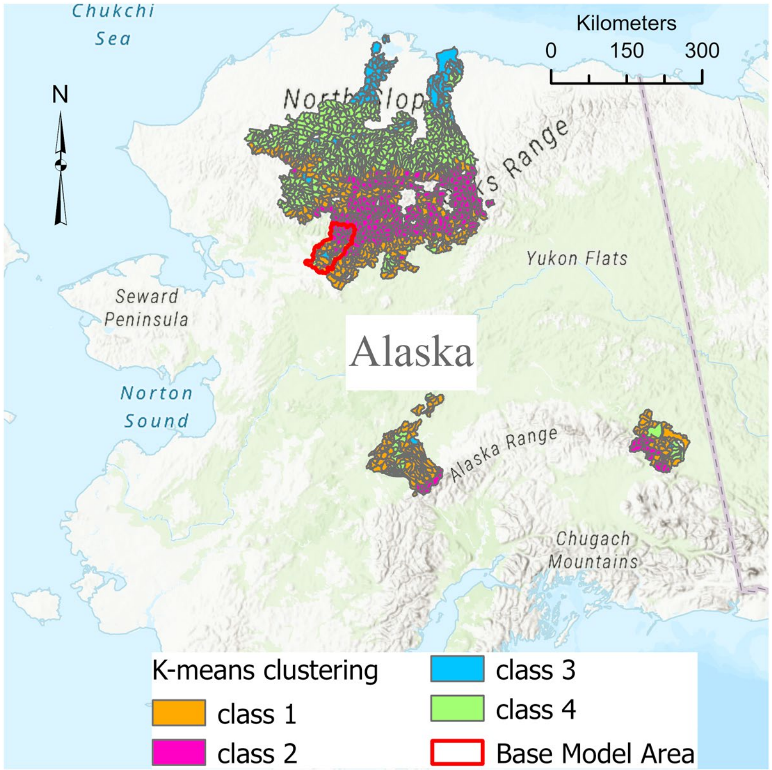

Figure 2: Distribution of four classes determined from k-means clustering of summary statistics for 12 data layers for the 12-digit hydrologic unit watershed in the study area domain.

An example of results from the k-means clustering for four classes is shown in Figure 2. Transfer learning results comparing tests of several cluster configurations will be presented at the conference and a review of the transfer model performance on proximal and distant watersheds.

This ongoing analysis will inform development of an efficient strategy for extending deep learning models across the varying geographic conditions in Alaska. Comparisons of the efficacy of retraining based on a systematic sample derived from attribute clustering with retraining based on a random sample will provide valuable information to deep learning modelers regarding appropriate retraining and model extension strategies.

Disclaimer: *Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.*

References

Cui, Y., Song, Y., Sun C., Howard, A., Belongie, S., 2018. Large scale fine-grained categorization of domain-specific transfer learning. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1806.06193.

Mitasova, H., Thaxton, C., Hofierka, J., McLaughlin, R., Moore, A., Mitas, L., 2004. Path sampling method for modeling overland waterflow, sediment transport, and short-term terrain evolution in Open Source GIS. Dev. Water Sci. 2004, 55, 1479–1490.

Pastick, N.J., Jorgenson, M.T., Wylie, B.K., Nield, S.J., Johnson, K.D., Finley, A.O., 2015. Distribution of near-surface permafrost in Alaska: Estimates of present and future conditions. Remote Sensing of Environment 168 (2015): 301-314. https://doi.org/10.1016/j.rse.2015.07.019.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A, Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J.B., Hofton, M., 2021. Mapping global forest canopy height, through integration of GEDI and Landsat data. https://doi.org/10.1016/j.rse.2020.112165.

Ronneberger, O., Fischer, P., Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland; New York, NY, USA.

Stanislawski, L.V., Jaroenchai, N., Wang, S., Shavers, E., Duffy, A., Thiem, P., Jiang, Z., Camerer, A., 2023. Transferring deep learning models for hydrographic feature extraction from IfSAR data in Alaska. GIScience 2023 Workshop on CartoAI: AI for cartography. Leeds, United Kingdom, Sept 12, 2023.

Stanislawski, L.V., Shavers, E.J., Wang, S., Jiang, Z., Usery, E.L., Moak, E., Duffy, A., Schott, J., 2021. Extensibility of U-net neural network model for hydrographic feature extraction and implications for hydrologic modeling. Remote Sensing. 2021; 13(12):2368. https://doi.org/10.3390/rs13122368.

Lawrence V. Stanislawski, Research Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO 65401

Ethan Shavers, Supervisory Geographer, U.S. Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO 65401

Neal Pastick, Research Physical Scientist, U.S. Geological Survey, Earth Resources Observation and Science Center, Sioux Falls, SD 57198

Phillip Thiem, Computer Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Science, Rolla, MO 65401

Shaowen Wang, Professor and Dean, Department of Geography & Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL 61801

Nattapon Jaroenchai, Graduate Student, Department of Geography & Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL 61801

Zhe Jiang, Assistant Professor, Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32611

Barry Kronenfeld, Professor, Geology and Geography Department, Eastern Illinois University, Charleston, IL 61920

Barbara P. Buttenfield, Professor Emerita, Department of Geography, University of Colorado-Boulder, Boulder, CO 80303

Adam Camerer, USGS Student Contractor, University of Missouri Science & Technology, Rolla MO 65401